CLAIMS

I/we claim:

[c1] A method in a computing system for selecting web pages, comprising:

defining a set of web page attributes;

for each of a plurality of web pages,

automatically extracting from the web page one or more attribute values, each for one of the set of attributes;

storing the extracted attribute values together with a URL for the web page in a dimensional model of the plurality of web pages, by adding a row for the web page to a fact table of the dimensional model, the added fact table row containing the URL and referencing, for each attribute for which an attribute value was extracted, a row corresponding to the attribute value in a dimension table of the dimensional model corresponding to the attribute;

receiving a query specifying query attribute values for one or more of the set of attributes;

processing the query against the dimensional model, by:

for each of the set of attributes for which a query attribute value is specified, selecting the rows of the dimension table corresponding to the attribute that match the query attribute value;

joining the selected rows of the dimension tables corresponding to the attributes for which a query attribute value is specified to the fact table to produce a join result; and

generating a query result containing the URLs contained by the rows of the join result.

[c2] The method of claim 1 wherein the generated query result contains one or more attribute values extracted from the web pages corresponding to the URLs contained by the rows of the join result.

[c3] The method of claim 1 wherein the plurality of web pages comprises web pages that are company home pages, and wherein the defined set of attributes includes a company name attribute, a company type attribute, a company category attribute, and a company location attribute.

[c4] The method of claim 1, further comprising identifying the plurality of web pages by:

initializing the plurality of web pages by contributing one or more known web pages to the plurality of web pages; and

reiteratively, from one of the plurality of web pages:

following a link on the web page to a new web page; and

contributing the new web page to the plurality of web pages.

[c5] The method of claim 1, further comprising, for each of the plurality of web pages, for each attribute for which an attribute value was extracted:

determining whether an existing row of the dimension table corresponding to the attribute corresponds to the attribute value;

if an existing row of the dimension table corresponding to the attribute corresponds to the attribute value, including in the row added to the fact table a reference to the existing row of the dimension table;

if no existing row of the dimension table corresponding to the attribute corresponds to the attribute value:

adding a new row corresponding to the attribute value to the dimension table; and

including in the row added to the fact table a reference to the new row added to the dimension table.

[c6] A method in a computing system for generating a dimensional model of a plurality of documents, comprising:

initializing a dimensional model comprising a fact table and a plurality of dimension tables;

for each of the plurality of documents,

obtaining one or more attribute values associated with the document, each for a document attribute;

adding a row for the document to the fact table;

storing in the row added to the fact table a reference to the document; and

for each attribute for which an attribute value was extracted, storing in the row added to the fact table a reference to a row corresponding to the attribute value in a dimension table corresponding to the attribute.

[c7] The method of claim 6, further comprising, for each of the plurality of documents, for each attribute for which an attribute value was extracted:

determining whether an existing row of the dimension table corresponding to the attribute corresponds to the attribute value;

if an existing row of the dimension table corresponding to the attribute corresponds to the attribute value, including in the row added to the fact table a reference to the existing row of the dimension table;

if no existing row of the dimension table corresponding to the attribute corresponds to the attribute value:

adding a new row corresponding to the attribute value to the dimension table; and

including in the row added to the fact table a reference to the new row added to the dimension table.

[c8] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a link to the referenced document.

[c9] The method of claim 6 wherein the reference to a document stored in at least one selected row of the fact table is a link to a list of links to documents.

[c10] The method of claim 9, further comprising, for each of the plurality of documents, storing in the row added to the fact table for the document an indication of the number of links to documents contained in the list of links to documents.

[c11] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is the title of the referenced document.

[c12] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a description of how to retrieve the referenced document.

[c13] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a link to a web page.

[c14] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a reference to a programmatic object.

[c15] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a reference to a Component Object Model component.

[c16] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a reference to a Common Object Request Broker Architecture component.

[c17] The method of claim 6 wherein the each reference to a document stored in a row of the fact table is a reference to a product description.

[c18] The method of claim 6 wherein attribute values are obtained for at least a portion of the plurality of documents by traversing links between documents of the plurality and parsing the contents of traversed-to documents.

[c19] The method of claim 6 wherein attribute values are obtained for at least a portion of the plurality of documents by receiving submissions indicating the attribute values associated with the document.

[c20] The method of claim 6, further comprising processing a document query using the dimensional model.

[c21] The method of claim 6 wherein the extracted attribute values for a selected one of the attributes are hierarchical, and wherein the dimension table row references added to the fact table for the selected attribute refer to dimension table rows corresponding to hierarchical attribute values.

[c22] A method in a computing system for collecting information about a group of documents, comprising:

seeding the group of documents with one or more known documents; and

for at least a portion of the documents of the group:

extracting from the document two or more attribute values describing the document, each of the extracted attribute values corresponding to a different document attribute;

identifying at least one link in the document to a linked-to document; and

adding to the group each linked-to document to which a link was identified.

[c23] The method of claim 22 wherein each of the documents added to the group is a web page.

[c24] The method of claim 22, further comprising, for each document:

applying a test to the document to determine whether to add the attribute values extracted from the document to a set of information collected about the group of documents; and

adding the attribute values extracted from the document to a set of information collected about the group of documents only if the applied test is satisfied.

[c25] The method of claim 22, further comprising, for each document:

applying a test to the document to determine whether to add linked-to documents to which a link was identified in the document to the group; and

adding the linked-to documents to the group only if the applied test is satisfied.

[c26] A method in a computer system for augmenting the data content of a document, comprising:

parsing the document comprised of semantically unstructured data to identify semantic attributes of the document; and

adding the identified semantic attributes to the document using semantically structured constructs.

[c27] A method in a computing system for processing a search request against a dimensional model of a set of documents, the model comprising a fact table and two or more dimension table, the fact table being comprised of rows each containing a document reference and referencing, for each attribute for which an attribute value was extracted, a row corresponding to the attribute value in a dimension table of the dimensional model corresponding to the attribute, the method comprising:

receiving a search request specifying search request attribute tests for one or more of the set of attributes;

for each of the set of attributes for which a search request attribute test is specified, selecting the rows of the dimension table corresponding to the attribute that satisfy the search request attribute test;

joining the selected rows of the dimension tables corresponding to the attributes for which a search request attribute value is specified to the fact table to produce a join result; and

generating a search request result containing the document references contained by the rows of the join result.

[c28] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds matches a value specified by the search request.

[c29] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds is non-null.

[c30] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds falls within a range specified by the search request.

[c31] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds matches a pattern specified by the search request.

[c32] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds is among a list of alternative values specified by the search request.

[c33] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds satisfies a programmatic function specified by the search request.

[c34] The method of claim 27, wherein one of the specified search request attribute tests tests whether the attribute to which it corresponds satisfies a mathematical function specified by the search request.

[c35] The method of claim 27, further comprising:

performing a word search upon the generated search result; and

generating a second search result conveying the results of the performed word search.

[c36] A method in a computer system for receiving a web page search request, comprising:

receiving the web page search request, the received search request specifying attribute values for each of one or more predetermined document attributes and requesting a report of web pages having the specified attribute values for those attributes; and

storing the received search request for processing.

[c37] The method of claim 36 wherein the search request is received from a source, the method further comprising transmitting a report of web pages having the specified attribute values for those attributes to the source.

[c38] A computing system for generating a dimensional model a plurality of documents, comprising:

a memory initially containing a dimensional model comprising a fact table and a plurality of dimension tables; and

a modeling subsystem that, for each of the plurality of documents,

obtains one or more attribute values associated with the document, each for a document attribute;

adds a row for the document to the fact table;

stores in the row added to the fact table a reference to the document; and

for each attribute for which an attribute value was extracted, stores in the row added to the fact table a reference to a row corresponding to the attribute value in a dimension table corresponding to the attribute.

[c39] The computing system of claim 38, further comprising a query execution subsystem that:

receives a search request specifying search request attribute values for one or more of the set of attributes;

for each of the set of attributes for which a search request attribute value is specified, selects the rows of the dimension table corresponding to the attribute that match the search request attribute value;

joins the selected rows of the dimension tables corresponding to the attributes for which a search request attribute value is specified to the fact table to produce a join result; and

generates a search request result containing the document references contained by the rows of the join result.

[c40] The method of claim 39 wherein the generated search request result contains, in conjunction with each document reference contained by the rows of the join result, additional attribute values for the referenced documents.

[c41] One or more computer memories collectively containing a dimensional document set model data structure, comprising:

a fact table comprising a set of rows, each fact table row corresponding to a document in a document set and containing a reference to the document; and

two or more dimension tables each comprising a set of rows, each dimension table corresponding to a different document attribute, the rows of each dimension table containing one of the unique values among the documents of the document set for the attribute to which the dimension table corresponds, wherein each fact table row further contains references to dimension table rows containing attribute values of the document to which the fact table row corresponds,

such that the contents of the data structure may be used to perform a document query expressed in terms of one or more attribute values.

[c42] The computer memories of claim 41 wherein each of the references to a document contained by the fact table rows is a link to the document.

[c43] The computer memories of claim 41 wherein each of the references to a document contained by the fact table rows is a link to a web page.

[c44] The computer memories of claim 41 wherein at least one of the references to a document contained by the fact table rows is a link to a list of links to documents having the same attribute values.

[c45] The computer memories of claim 44 wherein fact table rows containing references to a document that are links to a list of links to documents having the same attribute values further each contain an indication of the number of links in the list.

[c46] The computer memories of claim 44 wherein each of the references to a document contained by the fact table rows is a reference to a programmatic object.

[c47] The computer memories of claim 44 wherein each of the references to a document contained by the fact table rows is a reference to a Component Object Model component.

[c48] The computer memories of claim 44 wherein each of the references to a document contained by the fact table rows is a reference to a Common Object Request Broker Architecture component.

[c49] The computer memories of claim 44 wherein each of the references to a document contained by the fact table rows is a reference to a product description.

[c50] The computer memories of claim 35 wherein each of the references to a document contained by the fact table rows is a reference to instructions for obtaining a document.

[c51] One or more computer memories collectively containing an attribute-annotated web page data structure, comprising:

semantically unstructured content; and

two or more indications of semantic attribute values, each indication explicitly indicating an attribute and an attribute value, the explicitly-indicated attributes including at least two different attributes,

such that the web page may be modeled, indexed and/or searched on its semantic attribute values.

[c52] One or more generated data signals collectively conveying a dimensional document query data structure, comprising two or more indications of semantic attribute values, each indication explicitly indicating an attribute and an attribute value, the explicitly-indicated attributes including at least two different attributes,

such that the contents of the data structure may be used to select documents having the indicated attribute values.

[c53] A method in a computing system for selecting documents, comprising:

maintaining a dimensional model of a group of documents, the dimensional model reflecting values for a plurality of differentiated attributes for each of the documents of the group;

receiving a query specifying values for one or more of the plurality of attributes; and

in response to receiving the query, using the dimensional model to generate a list of documents in the group having the attribute values specified by the query.

[c54] The method of claim 53 wherein maintaining the dimensional model includes automatically extracting attribute values from the documents.

[c55] The method of claim 54 wherein attribute values are extracted from one or more explicit attribute tags within the documents.